Estimation de la loi de probabilité de l'âge au décès
Méthode bayésienne

Notice d'utilisation des programmes en R.

Henri CAUSSINUS et Daniel COURGEAU

Avertissement. *Les programmes présentés sont des outils de travail qui, dans leur état actuel, ne prétendent pas utiliser au mieux toutes les ressources du logiciel R. Ils convient donc de les considérer comme provisoires et susceptibles de nombreuses améliorations.*

Le programme **Estimation1** considère une loi a priori de Dirichlet pour les probabilités des classes d'âge.

Le programme **Estimation2** considère deux lois a priori : Dirichlet comme le premier et loi uniforme sur un ensemble de vecteurs candidats. La notice ci-après considère d'abord le programme **Estimation1.** Elle donne ensuite les modifications concernant l'utilisation du programme **Estimation2**.

Estimation1

Objectif et résultats fournis

Partant de la répartition de stades osseux sur un site cible et de données de référence, le programme estime les probabilités des différentes classes d'âge selon la méthode bayésienne préconisée par Caussinus et Courgeau dans le chapitre B du Manuel de Paléodémographie (voir aussi Caussinus et Courgeau, *Population*, 2010). La méthode est non paramétrique en ce sens qu'elle n'utilise pas de modèle de mortalité explicite. On peut cependant compléter les calculs par l'ajustement d'une loi de mortalité (Gompertz ou Gompertz-Makeham).

Le programme fournit :

1) Des paramètres de la loi a posteriori : moyennes, écarts-types, matrice des coefficients de corrélation, quantiles de chaque loi marginale.

2) Des graphiques des fonctions de répartition a posteriori pour chaque classe d'âges, ainsi que des densités a posteriori.

3) L'ajustement d'une loi de Gompertz et d'une loi de Gompertz-Makeham : estimation des paramètres et représentation graphique.

Données statistiques

Les âges et les stades osseux sont répartis en classes. On dispose des fréquences des divers stades osseux sur un site ainsi que de données de référence sous forme d'une matrice de fréquences (stades en ligne, âges en colonne).

Les données de référence sont à lire dans un fichier texte (séparateur = blanc). On trouvera en complément des programmes un certain nombre de ces fichiers issus de la population de référence de Lisbonne pour différents nombres de stades et d'âges, et selon le/les sexe(s) considérés (par exemple Lisb5x8H pour 5 stades, 8 groupes d'âge et pour les hommes seuls). Il est possible de reconstituer n'importe quelle autre table à partir de ces données détaillées fournies dans ce CD-ROM (fichiers.txt).

Mettre le nom du fichier dans la commande

ref<-read.table("nom du fichier.txt").

Les données de site sont à introduire au clavier sous la forme :

sit<- c(suite des fréquences avec un point pour les décimales et une virgule pour séparateur)

Il faut enfin introduire les limites des classes d'âge sous la forme :

a0<- limite inférieure (en années)

ac<-c(limites suivantes en années séparées par une virgule)

Attention : pas de limite supérieure pour la dernière classe qui est supposée ouverte.

(Remarque : les limites de classe n'interviennent pas dans les calculs sauf pour l'ajustement d'une loi de mortalité, mais elles sont nécessaires pour la présentation des résultats).

Données de réglage

Les paramètres de la loi a priori de Dirichlet sont à introduire sous la forme :

bet<-c(suite des valeurs avec un point pour les décimales et une virgule pour séparateur)

Pour les autres données, on peut modifier les valeurs par défaut du programme mais ce sera en général peu utile.

Il s'agit de :

ns = nombre de répétitions dans l'évaluation des intégrales par la méthode de Monte Carlo (par défaut 100000).

coeff = coefficient de réduction appliqué aux données de référence (par défaut 1).

na = nombre de points d'abscisse dans la discrétisation des fonctions de répartition a posteriori (par défaut = 1000).

nad = nombre de points d'abscisse dans la discrétisation des densités a posteriori (par défaut = 100 ; nad inférieur à na est préconisé car les densités sont utilisées pour des représentations graphiques et il semble avant tout important d'en donner une version suffisamment lissée).

q = vecteur des probabilités correspondant aux quantiles que l'on veut exprimer (par défaut q= c(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)).

init = valeurs initiales pour le processus itératif utilisé dans l'ajustement d'une loi de Gompertz (par défaut init = c(0.0086, 0.067)).

Structure du programme

Le programme est organisé de la façon suivante :

- lecture / introduction des données et paramètres de réglage ; vérification à la console.
- estimation de la loi a posteriori - calculs de base.
- représentations graphiques des densités a priori et a posteriori.
- ajustement d'une loi de Gompertz et d'une loi de Gompertz-Makeham.

Résultats

Les résultats numériques fournis directement sont :

- Moyennes et écarts-types de la loi a posteriori pour chaque classe d'âges,
- Matrice des coefficients de corrélation de la loi a posteriori,
- Quantiles de la distribution a posteriori pour chaque classe d'âges.

Ces résultats sont donnés sur la console et sont aussi enregistrés dans un fichier texte nommé « résultats ».

Les représentations graphiques comparent, pour chaque classe d'âges, les densités a priori (rouge) et des densités a posteriori (noir) ; sur ces graphiques sont placés des traits verticaux dont les abscisses sont les moyennes a priori (rouge) et a posteriori (noir).

Par défaut, le programme affiche les graphiques correspondant à 9 classes d'âges au maximum. Le graphique affiché peut être enregistré. Pour afficher moins de classes ensemble (ou davantage mais ce n'est pas conseillé) il convient de modifier le « layout » en :

layout (matrix(1:nombre de clases affichées, nombre de lignes, nombre de colonnes, byrow=TRUE))

Pour afficher les classes de m à n, remplacer dans la ligne qui suit (h in 1:7) par (h in m:n) ; pour afficher des classes de numéros quelconque, par exemple les classes 1,4,6,8,9, remplacer (h in 1:7) par (h in c(1,4,6,8,9)).

Les figures peuvent être enregistrées : dans le menu Fichier, faire « sauver sous » et choisir le mode de sauvegarde. Noter qu'il sera souvent préférable d'enregistrer une figure après avoir agrandi la fenêtre dans laquelle elle est présentée.

Attention :

- ne pas inclure un numéro de classe supérieur au nombre de classes de l'exemple traité,
- les figures doivent être visualisées (et enregistrées si on le souhaite) une à une, ce qui implique d'activer le programme étape par étape dans la production de graphiques.

Le programme permet d'ajuster une loi de Gompertz et une loi de Gompertz-Makeham aux probabilités de décès par âges estimées (voir Annexe B). Il fournit :

- les valeurs estimées des paramètres de ces lois, ainsi que l'écart quadratique résiduel,
- un graphique comparant les probabilités de survie estimées par la méthode bayésienne (ronds pour chaque limite de classes), les probabilités de survie selon la loi de Gompertz ajustée (courbe rouge), les probabilités de survie selon la loi de Gompertz-Makeham ajustée (courbe verte).

Estimation2

Le programme **Estimation2** fournit les mêmes résultats que le programme **Estimation1** pour une loi a priori de Dirichlet, plus les résultats analogues avec une loi a priori uniforme sur un ensemble de vecteurs « candidats ». Cet ensemble est inspiré de la méthode de Bocquet-Appel et Bacro ; nous préconisons d'utiliser les ensembles fournis sur le site de Bocquet-Appel lorsque les conditions de l'application le permettent. En outre, le programme donne quelques éléments de comparaison des deux façons d'utiliser la méthode bayésienne préconisée. Nous indiquons seulement ci-après les modifications qui interviennent par rapport au programme **Estimation1**. On trouvera quelques éléments méthodologiques sur le choix de la loi a priori dans l'annexe C.

Données

Les vecteurs candidats viennent d'un fichier texte (séparateur = blanc) et sont chargés par la commande :

vp<-read.table("nom du fichier.txt")

(le programme est conçu afin d'utiliser directement les fichiers disponibles sur le site de Bocquet-Appel : c'est le sens de l'instruction vp<-t(vp[,1:7]) qui sera à modifier pour un autre type de fichier).

Résultats

Ils sont donnés selon les mêmes principes que pour **Estimation1** (console et enregistrement dans le fichier « résultats », graphiques) mais doublés puisque les résultats correspondant à chacune des lois a priori sont fournis. En, outre, on peut obtenir des graphiques comparatifs : comparaison par classes d'âges des deux densités a priori et des deux densités a posteriori.

Remarque. La loi a priori uniforme sur un nombre fini de vecteurs et, consécutivement, la loi a posteriori correspondante sont des lois discrètes : les densités n'ont plus le même sens que pour les précédentes lois (continues) ; ce ne sont donc pas exactement les densités que donnent les graphiques, mais des courbes qui s'en veulent l'expression visuelle (lissée).

Annexe A. Le logiciel R : principes d'utilisation

R est un logiciel libre qui peut être téléchargé à partir de plusieurs sites, par exemple
http://cran.cict.fr

R peut être utilisé sous Windows, MacOS X ou Linux. Voici comment procéder pour utiliser sous Windows un des programmes décrits ici (pour l'utilisation sous MacOS X ou Linux se reporter au site précédent).

- Télécharger d'abord R.2.10.1 pour Windows à partir du site vers le dossier de fichiers *Program Files*, qui crée un dossier nommé R et met simultanément en place une icône R sur le bureau et le menu démarrer.
- Créer ensuite un dossier qui sera le "répertoire courant" et dans lequel on mettra les programmes et les fichiers utilisés (en particulier les fichiers des données de référence pour tous les programmes et les fichiers de vecteurs candidats pour Estimation2).
- Cliquer sur l'icône R. Cela fait apparaître la « console ».
- Aller dans le menu Fichier et faire "changer le répertoire courant" ce qui permet de se placer dans le répertoire préparé à cet effet.
- Aller de nouveau dans le menu Fichier et faire "ouvrir un script" : on obtient le répertoire dans lequel se trouvent les programmes et appelle le programme désiré. Une nouvelle fenêtre est alors ouverte qui contient ce programme.
- Pour actionner une partie du programme, il faut la sélectionner à la souris et cliquer sur la troisième icone en haut à gauche (en forme de fenêtre avec une flèche au milieu). Cela active la partie sélectionnée du programme ; celle-ci est déroulée dans la console, qui affiche aussi les résultats obtenus si de tels résultats sont prévus dans cette partie du programme.

Remarques

Dans le programme, ce qui est mis entre apostrophes 'comme cela' n'est pas exécuté ; il s'agit de commentaires.

Dans un premier temps, il est conseillé d'exécuter le programme partiellement en s'aidant des commentaires ; par exemple commencer en allant jusqu'à 'vérification des données' (elles apparaissent sur la console), etc. Une telle exécution partielle est indispensable pour les graphiques (voir plus haut).

Les résultats sont lus à la console mais aussi enregistrés dans un fichier « résultats » si on le souhaite (en activant la partie correspondante du programme). Ce fichier se trouvera dans le répertoire courant ; on peut le lire (et l'extraire) en allant dans ce répertoire (faire toujours "Fichier" pour cela) mais il faut alors faire venir "all the files" car, en principe, par défaut, seuls les programmes apparaissent. Un point important : le fichier « résultats » est conçu pour contenir la suite des résultats correspondant à une application donnée, et seulement à celle-ci ; une nouvelle application efface donc les résultats précédemment enregistrés ; le plus simple est de les conserver dans un fichier portant un autre nom ou situé ailleurs que dans le répertoire courant.

Annexe B. Estimation de la loi de Gompertz ou de Gompertz-Makeham la plus proche de la distribution observée

1. La loi de Gompertz

En 1825, Gompertz s'appuyant sur des données de divers pays, énonça la loi qui porte désormais son nom et qui est souvent applicable à la mortalité des adultes, au moins jusqu'à 80 ans. Elle représente le quotient instantané de mortalité par la formule :

$$h(x) = -\frac{1}{S(x)}\frac{\partial S(x)}{\partial x} = \lambda\rho e^{\rho x}$$

où $S(x)$ est la probabilité de survivre de la naissance à l'âge x, λ et ρ deux paramètres à estimer. Le paramètre λ pose le niveau général de la mortalité et le paramètre ρ mesure la réduction de la capacité biologique des individus à lutter contre la mort lorsque l'âge augmente. On peut intégrer cette équation différentielle, qui donne :

$$\ln[S(x)] = -\int_{a=0}^{x} \lambda\rho e^{\rho a}\, da = \lambda\left(1 - e^{\rho x}\right).$$

On peut également écrire le quotient cumulé de mortalité entre deux âges x et $x + n$:

$${}_nH_x = \int_{\xi=x}^{x+n} \lambda\ \rho e^{\rho\xi} d\xi = \lambda\left(e^{\rho(x+n)} - e^{\rho x}\right) = \lambda e^{\rho x}\left(e^{\rho n} - 1\right).$$

On voit donc que ce quotient cumulé suit une loi exponentielle de l'âge x comme le quotient instantané mais avec un coefficient λ différent, le coefficient ρ restant le même. Mais ce quotient cumulé est différent du quotient démographique pluriannuel classique.

On peut passer des quotients instantanés aux quotients pluriannuels par la formule :

$${}_nq_x = 1 - \exp\left(-\int_{\zeta=x}^{x+n} h(\zeta)\, d\zeta\right)$$

où h est le quotient instantané.

Avec une loi de Gompertz, cela donne :

$${}_nq_x = 1 - \exp\left(-\int_{\xi=x}^{x+n} \lambda\rho e^{\rho\xi} d\xi\right) = 1 - \exp\left(\lambda\left\{e^{\rho x} - e^{\rho(x+n)}\right\}\right).$$

Il en résulte que l'on peut écrire la probabilité, à l'âge x, de survivre au moins jusqu'à l'âge $x+n$:

$${}_np_x = 1 - {}_nq_x = \exp\left(\lambda\left\{e^{\rho x} - e^{\rho(x+n)}\right\}\right).$$

On voit donc que c'est le logarithme népérien de cette probabilité de survie qui suit également une loi de Gompertz, la même que celle du quotient cumulé.

En 1860, Makeham considère qu'il faut tenir compte du fait que certaines causes de décès sont indépendantes de l'âge, telles que les accidents, et propose de compléter la loi de Gompertz par un coefficient constant λ_0 :

$h(x) = \lambda_0 + \lambda\rho e^{\rho x}$.

On voit que dans ce cas en calculant la probabilité, à l'âge *x*, de survivre au moins jusqu'à l'âge *x*+*n* :

${}_n p_x = \exp\left(\lambda_0 n + \lambda e^{\rho x}\left\{e^{\rho n} - 1\right\}\right)$,

le logarithme de cette probabilité de survie suit également une loi de Gompertz-Makeham, de paramètres différents de la précédente.

2. Estimation

La méthode bayésienne donne une estimation des probabilités des diverses classes d'âge, par une loi a posteriori. Soit μ le vecteur (colonne) des moyennes des probabilités de survie estimées et V la matrice des variances et covariances de cette loi.

Soit $G(\lambda,\rho)$ le vecteur (colonne) des probabilités des classes sous l'hypothèse d'une loi de Gompertz de paramètres λ et ρ (et conditionnellement à la survie au-delà du plus petit des âges considérés).

Une façon simple d'ajuster une loi de Gompertz, c'est-à-dire de déterminer les paramètres λ et ρ les mieux compatibles avec les observations, est de procéder par moindres carrés en minimisant en λ et ρ l'expression :

$$(\mu - G(\lambda, \rho))' (\mu - G(\lambda, \rho)) \qquad (1)$$

Il peut cependant paraître plus naturel d'utiliser des moindres carrés convenablement pondérés et de minimiser en λ et ρ l'expression :

$$(\mu - G(\lambda, \rho))' V^{-1} (\mu - G(\lambda, \rho)) \qquad (2)$$

Dans la pratique il faut omettre une classe dans *G* pour que V ne soit pas nécessairement singulière ; par ailleurs, nous passons plutôt par les probabilités cumulées afin de simplifier l'expression de *G* en faisant les transformations convenables de μ et V (cela ne change pas l'estimation de λ et ρ en vertu de la linéarité de cette transformation). On a alors pour les c-1 éléments du vecteur $G(\lambda, \rho) = \exp\left(\lambda\left\{e^{\rho x_0} - e^{\rho x}\right\}\right)$ où x_0 est la limite inférieure de la première classe d'âge et *x* prend successivement la valeur des c-1 limites inférieures des classes d'âges suivantes.

Pour ajuster une loi de Gompertz-Makeham, on procède de façon analogue en remplaçant *G* (λ, ρ) par le vecteur des probabilités des classes sous l'hypothèse de cette nouvelle loi $G(\lambda_0, \lambda, \rho) = \exp\left(\lambda_0\{x - x_0\} + \lambda\left\{e^{\rho x_0} - e^{\rho x}\right\}\right)$.

Le minimum de (1) ou de (2) s'obtient facilement par une procédure standard. Les programmes utilisent la procédure constrOptim de R afin d'imposer la contrainte de positivité de λ et ρ. Une seule difficulté s'est présentée dans la pratique : si la réduction à c-1 éléments indiquée plus haut permet bien d'éviter que V soit singulière, cette matrice peut se trouver très proche de la singularité dans le cas de la loi a priori uniforme sur un ensemble de vecteurs, ce qui conduit à des résultats aberrants quand on cherche à minimiser (2). Pour cette raison, le programme **Estimation2** utilise par défaut la minimisation de (1) dans le cas de la loi a priori uniforme. On peut cependant lui demander facilement de minimiser (2) : il suffit de mettre entre apostrophes l'instruction 'vinvU<-diag(nc-1)' à la ligne 11 de la partie 'lois de Gompertz et Gompertz-Makeham ajustées' du programme.

Annexe C. Sur le choix de la loi a priori.

Quelques éléments de discussion sur le choix des lois a priori sont donnés dans la chapitre B du Manuel de Paléodémographie et dans l'article de *Population* (Caussinus et Courgeau, 2010). On y trouve en particulier des arguments pour le choix des paramètres de la loi a priori de Dirichlet. Nous ne reviendrons pas ici sur ces aspects. Nous les complétons par quelques considérations sur l'alternative : loi de Dirichlet - loi uniforme sur un ensemble de vecteurs.

Nous avons montré dans le chapitre B du Manuel que la méthode bayésienne avec loi de Dirichlet convenable domine les anciennes méthodes d'estimation ; dans l'article de *Population*, nous notons que la seule méthode non paramétrique susceptible de concurrencer la nôtre est celle de Bocquet-Appel et Bacro ; nous comparons donc notre méthode à celle de ces auteurs en supposant que le vecteur de probabilité à estimer est un de leurs vecteurs « candidats ». Nous montrons alors que la façon de procéder la plus efficace est d'utiliser la méthode bayésienne avec loi a priori uniforme sur ces vecteurs. Cela conforte notre méthode par rapport à celle de Bocquet-Appel et Bacro, puisque nous nous sommes placés dans le cas qui semblait a priori le plus favorable à cette dernière ; par ailleurs, cela donne un avantage à la loi a priori uniforme sur la loi de Dirichlet (avec paramètres déduits du standard préindustriel) lorsque le vecteur à estimer est l'un des vecteurs utilisés pour la loi a priori uniforme. Or, si l'on regarde de près les implications d'une telle loi sur les lois a priori des probabilités de chacune des classes d'âges, on voit que celles-ci ont une dispersion relativement faible comparée à la dispersion des lois a priori de Dirichlet[1]. Si l'on est donc en mesure d'affirmer que la connaissance « a priori » du site considéré est suffisamment précise et compatible avec la loi a priori uniforme, celle-ci est raisonnable ; sinon, nous pensons qu'il vaut mieux s'en tenir à une loi de Dirichlet dont l'impact sur l'estimation finale est bien moindre. En d'autres termes, si les lois a posteriori sont en général moins dispersées avec a priori uniforme qu'avec a priori de Dirichlet, on doit craindre qu'il s'agisse d'une précision trompeuse, conséquence d'une loi a priori injustifiée. Les graphiques inclus dans le programme Estimation2 comparant les deux types de lois a priori et a posteriori peuvent éclairer l'utilisateur.

[1] A titre d'exemple, avec le fichier de vecteurs candidats "ProbAtri20-90.txt" la loi a priori de la classe 50-59 ans est concentrée entre 0,112 et 0,233 ce qui interdit une estimation ponctuelle sortant de cet intervalle (aussi bien par notre méthode avec a priori uniforme que par la méthode de Bocquet-Appel et Bacro).